ZeRCP: Towards Communication-Efficient Collaborative Perception and Future Scene Prediction via Request-Free Spatial Filtering

Yijie Chen[1], **Yuzhe Ji**[1], **Haotian Wang**[1], **Xiaoyun Qiu**[1], **Yingcong Chen**[1,2], **Xinhu Zheng**[1,2*]

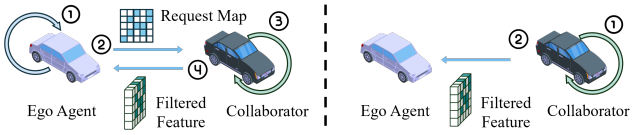
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]The Hong Kong University of Science and Technology
{ychen324, yji755, xqiu329}@connect.hkust-gz.edu.cn, {haotianwang, yingcongchen, xinhuzheng}@hkust-gz.edu.cn

Abstract

Multi-Agent collaboration addresses inherent limitations of individual agent systems, including limited sensing range and occlusion-induced blind spots. Despite significant progress, persistent challenges such as constrained communication bandwidth and under-explored subsequent extensions still hinder real-time deployment and further developments of collaborative autonomous driving systems. In this work, we propose ZeRCP, a unified communication-efficient framework that bridges collaborative perception with future scene prediction. Specifically, (i) we devise a plug-and-play request-free spatial filtering module (ZeroR) that eliminates the reliance on request maps while preserving inter-agent spatial complementarity modeling. This approach further reduce communication latency and bandwidth consumptions. (ii) We design a multi-scale pyramidal prediction network anchored by a novel Spatial-Temporal Deformable Attention (STDA) module, extending frame-wise detection to multi-frame predictions. This method adeptly models spatiotemporal dynamics without relying on auto-regressive recursion. We evaluate our method on a large-scale dataset in challenging semantic segmentation and scene prediction tasks. Extensive experiments demonstrate the superiority and effectiveness of ZeRCP in bandwidth-constrained collaboration scenarios and spatiotemporal prediction applications.

Introduction

Autonomous Driving (AD) emerges as a transformative technology to enhance driving experience and boost traffic efficiency. With the significant advancements in deep learning, single-agent perception systems have achieved remarkable progress (Lang et al. 2019; Liu et al. 2021). However, such individual systems remain constrained by several inherent limitations, particularly restricted sensing ranges and occlusion-induced blind zones. Recently, multi-agent collaboration (Xu et al. 2022c; Yu et al. 2022; Yuan et al. 2021) has gained considerable attention as a promising solution to these limitations. Collaboration systems facilitates data sharing among connected agents via Vehicle-to-Vehicle/Everything (V2V/X) communication, thereby deriving comprehensive perception results. Despite these approaches achieve substantial improvements in perception accuracy (Xu et al. 2024) and robustness (Li et al. 2023), communication efficiency persists as an outstanding challenge, especially in real-world deployments.



(a) Dual-Phase Protocol (b) Single-Phase Protocol (Ours)

Figure 1: The comparison of two protocols. Single-phase reduces bandwidth requirements and communication latency.

To address this critical bottleneck, several notable studies establish primary trade-off between perception performance and communication bandwidth via feature compression over channel-wise (Li et al. 2021) and spatial-wise (Yang et al. 2023a). Most spatial compression methods follow a request-triggered spatial filtering paradigm pioneered in Where2Comm(Hu et al. 2022b). In this paradigm, the ego agent first transmits a request map to collaborators, who then interact with this map to perform spatial filtering before returning collaborative information. We categorize these approaches as *dual-phase communication protocol*. While effective for compression, this approach inherently exacerbates temporal asynchrony, which is a non-trivial issue for such delay-sensitive systems. In contrast, *single-phase communication protocol* enables collaborators to locally perform spatial filtering without request maps, thereby reducing communication latency and bandwidth requirements. Figure 1 provides a concise illustration of these two protocols. However, existing single-phase methods primarily emphasize on foreground objects, overlooking inter-agent spatial complementarity and semantic information (Wang et al. 2023a; Zhao, ZHANG, and Zou 2023), thereby leading to suboptimal performance.

Furthermore, although several existing studies attempt to incorporate short-term historical cues to mitigate temporal asynchrony and improve perception accuracy (Wei et al. 2023; Yu et al. 2023; Li et al. 2025), most of these works remain largely confined to frame-by-frame detection, not extending to downstream tasks such as future prediction and ultimate planning. Consequently, this limitation hinders fur-

*Corresponding Author

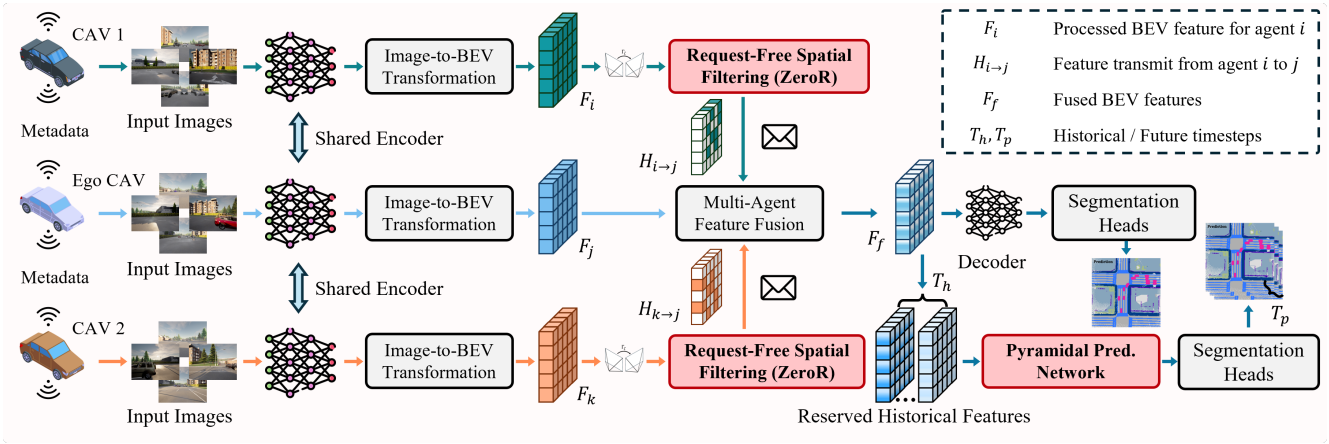


Figure 2: Overview of the proposed ZeRCP framework, comprising four core stage. Connected agents first exchange metadata for collaboration preparation. Subsequently, image features are extracted via a shared backbone and transformed to BEV representation. Next, our ZeroR module actively filters perceptual critical and complementary features for transmission and fusion. Eventually, the fused BEV features F_f enables downstream tasks: (i) real-time BEV semantic segmentation via specific decoder and heads, (ii) future scene prediction through a multi-scale pyramidal prediction network anchored by STDA module.

ther developments and applications of integrated collaborative AD systems. Recently, several preliminary frameworks emerge for cooperative trajectory prediction (Zhou et al. 2024; Ren, Chen, and Zhang 2024), yet extensions to subsequent tasks remain predominantly unexplored.

Building on these observations, we propose ZeRCP, a unified communication-efficient framework that bridges collaborative perception with future scene prediction, as shown in Figure 2. It integrates two core components: a plug-and-play spatial compression module ZeroR that operates under a single-phase communication protocol without any requests; and a multi-scale pyramidal prediction network centered on a novel STDA module that extends frame-wise detection to multi-frame prediction. In specific, (i) ZeroR enables collaborators to proactively filter perceptually critical and spatially complementary feature cells with the guidance of relative pose. Moreover, it employs a deformable attention-based spatial merging component, which concurrently expands receptive fields and disambiguates BEV spatial correlations compromised by monocular depth ambiguities. (ii) STDA addresses the core challenge of future prediction in autonomous driving: modeling spatiotemporal scene dynamics. We devise an EvoFlow block to capture egocentric scene evolutions patterns. Subsequently, a 3D deformable attention component jointly models spatiotemporal correlations by implicitly encoding them into learnable 3D offsets (t, x, y). This method achieves multi-frame prediction without auto-regressive recursion structure and any auxiliary inputs beyond historical intermediate features.

To validate the effectiveness of our framework, we employ multi-class semantic segmentation as our target task. This task presents more challenging for spatial compression under limited bandwidth compared to the widely adopted object detection task. Similarly, as a natural extension of perception, we implement multi-frame semantic scene forecasting, which provides well-suited alignment with perception pipelines while exhibiting more comprehensive scene understanding. We evaluate our framework on the public large-scale collaborative perception dataset OPV2V. Comprehensive experimental results demonstrate ZeroR's superiority in communication-efficient collaboration and STDA's effectiveness in spatiotemporal prediction applications. The main contributions can be summarized as follows:

- We develop a plug-and-play communication module ZeroR under the single-phase protocol. This design avoids extra request-map roundtrip while preserving inter-agent spatial complementarity; expands receptive fields and repairs BEV spatial correlations during filtering; and eventually enables communication-efficient and high-performance collaboration.
- We introduce a pyramidal prediction network centered on STDA module to extend frame-wise detection to multi-frame prediction. By leveraging learnable 3D offsets to jointly model spatiotemporal dynamics, STDA provides a novel paradigm for sequence prediction schemes.
- Our framework is evaluated on the large-scale OPV2V dataset towards challenging semantic-related tasks. Experimental results demonstrate the superiority and effectiveness of proposed ZeRCP in bandwidth-constrained collaboration and spatiotemporal prediction applications.

Related Work

Communication Efficiency

Channel-Wise Compression In recent years, significant research efforts have been directed towards reducing transmission volumes within the intermediate fusion process. Building upon this principle, many works apply channel-wise compression techniques to intermediate BEV features, serving as a computationally tractable paradigm to investigate the performance-communication trade-off in collaborative systems (Li et al. 2021; Xu et al. 2022b). The

core method lies in utilizing 1×1 convolutional layers, while CodeFilling (Hu et al. 2024) introduces a code-book that achieves pragmatic approximation of features along the channel dimension.

Spatial-Wise Compression On the other hand, emerging studies investigate the potential of spatial-wise feature compression. The pioneer work Where2Comm (Hu et al. 2022b) dynamically selects perceptually critical regions via inter-agent confidence map interactions. Following this request-triggered paradigm, How2Comm (Yang et al. 2023a), What2Comm (Yang et al. 2023c) and UMC (Wang et al. 2023b) continue to enhance the trade-off between perception performance and communication efficiency.

From the other perspective, CORE (Wang et al. 2023a) straightly samples critical feature cells with the highest top-k activation values without inter-agent coordination. Meanwhile, CoCa3D (Hu et al. 2023) employs static thresholding on confidence maps to focus on foreground features. These push-based approaches belong to *single-phase communication*, which our method is formulated within. Moreover, compared to existing single-phase approaches, our method implements a content-adaptive spatial selection mechanism.

Probabilistic Future Prediction

Individual Prediction In individual autonomous driving, early predictive capabilities focused on trajectory. Recently, researcher shift towards more holistic environment representation, such as occupancy forecasting. Frameworks like ConvLSTM (Shi et al. 2015) demonstrate great adaptability for such tasks. Additionally, some multi-modal fusion frameworks further enhance prediction accuracy (Hu et al. 2022c; Liu, Huang, and Lv 2023; Zyrianov et al. 2025). This occupancy-centric formulation tend to abstract away fine-grained semantic categories and these methods exhibit significant dependency on auxiliary inputs, such as HD maps, dynamic trajectories, and flow data. In end-to-end autonomous driving schemes, FIERY (Hu et al. 2021) is the first work to address probabilistic future prediction on BEV space with surrounding images input. Building on this method, ST-P3(Hu et al. 2022a) devises a dual pathway strategy to further enhance temporal modeling. Our method diverges from this paradigm, opting not to rely on auto-regressive recursion-based prediction.

Collaborative Prediction In the realm of multi-agent collaboration, several frameworks incorporate short-term historical cues to mitigate asynchrony and improve detection accuracy (Yu et al. 2023; Wei et al. 2023; Yang et al. 2023b). Nonetheless, they do not extend their focus to subsequent tasks until recently, when some efforts concentrate on trajectory prediction. For instance, CMP (Wang et al. 2025) adheres to detection-then-tracking pipelines and explores trajectory aggregation across multiple agents, while V2XPnP (Zhou et al. 2024) proposes a unified Transformer-based architecture for multi-agent spatiotemporal feature fusion. Despite these advancements, a more suitable and efficient multi-agent prediction pipeline remains lacking.

Methodology

Framework Design

The overall architecture of ZeRCP is illustrated in Figure 2. Given multi-view camera input $I_i \in \mathbb{R}^{V\times 3\times H_{img}\times W_{img}}$ from each agent $i \in \mathcal{A}$, image features are extracted and transformed onto the unified BEV representation through shared backbones. Following the work (Xu et al. 2022a), we employ the SinBEVT, with fused axial attention (FAX) as the core component, to perform this BEV feature computation process. Then, these BEV features $F_i \in \mathbb{R}^{C\times H\times W}$ are geometrically warped onto the ego's coordinate system via a differentiable spatial transformation operator Γ_ζ, deriving $H_i = \Gamma_\zeta(F_i) \in \mathbb{R}^{C\times H\times W}$.

Consequently, the ZeroR module actively selects perceptual critical and spatial complementary features $H_{i\to ego}$ for transmission. Upon receiving features from collaborators, a fusion operator is employed to aggregate multi-agent features into a global feature representation F_f. Notably, the fusion methodology is not our primary focus, thus, our framework can accommodate various existing fusion methods for comprehensive evaluation during experiments.

Subsequently, our multi-task decoding structure branches as follows: frame-wise segmentation stream directly decodes fused features to target BEV resolution through a CVT decoder (Zhou and Krähenbühl 2022), while the prediction pathway concatenates fused features with preserved historical features before processing through the prediction network. Eventually, we leverage the same yet not weight-shared head to achieve target tasks.

Base Model

As demonstrated in (Li et al. 2022), deformable attention mechanisms can adaptively sample features from data-dependent spatial locations rather than relying on fixed geometric patterns. In specific, it adaptively shifts reference points toward regions of interest, enabling to attend to semantically critical regions and model inter-region relevant. This mechanisms can be formulated as Equation 1.

$$\begin{aligned}\mathrm{DeformAtt}(q,v,p) &= \sum_{i=1}^{N_h}\mathcal{W}_i\sum_{j=1}^{N_k}\mathcal{A}_{ij}\cdot\mathcal{W}'_i\cdot v(p+\Delta p_{ij})\\ \text{where}\quad \mathcal{A} &= \mathrm{softmax}(\mathcal{W}_\mathcal{A}\cdot(V||Q)+b_\mathcal{A})\\ \Delta p &= \mathcal{W}_p\cdot(V||Q)+b_p\end{aligned} \tag{1}$$

where q, v, p represent the query, value and reference point, respectively. Let i and j index attention heads and sampled keys respectively, thus, N_h and N_k denotes the number of attention heads and the number of sampled key per head. $\mathcal{W}_i$ and $\mathcal{W}'_i$ are learnable weights for multiple heads. $\mathcal{A}_{ij} \in [0, 1]$ is the predicted attention weight for certain sampled key, and have been normalized by $\mathrm{softmax}(\cdot)$ operation. $v(p + \Delta p_{ij})$ represents the sampled feature at location $p + \Delta p_{ij}$, where $\Delta p_{ij} \in \mathbb{R}^n$ is the predicted offsets, having the same n dimension with p. Here, attention weights $\mathcal{A}$ and offsets Δp are derived from query q and value v through projection operators.

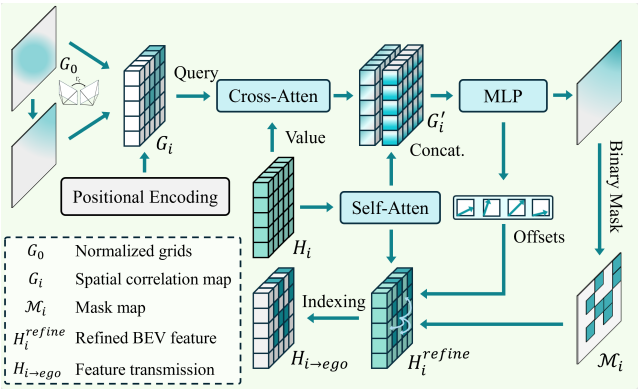


Figure 3: The architecture of the proposed ZeroR module that actively filters perceptual critical and spatial complementary information.

Request-Free Spatial Filtering (ZeroR)

In this section, we introduce a spatial filtering method that operates without any request or inter-agent feature interactions, relying exclusively on their local feature distributions and relative poses established in mete-data stage. We believe that the relative pose is an effective guidance to model inter-agent spatial correlation (Huang et al. 2024).

Spatial Correlation Modeling To begin with, a transformation matrix that can efficiently represent their relative poses is computed. Subsequently, we simulate coordinate transformations on a normalized grid $G_0 \in \mathbb{R}^{H\times W}$, and compute distance vector field with original ones, yielding a spatial correlation feature. Integrated with a 2D Sinusoidal Positional Encoding (Vaswani et al. 2017), we derive an initial spatial correlation map $G \in \mathbb{R}^{C_{emb}\times H\times W}$.

Attentive Spatial Filtering Given the initial correlation map G and BEV feature H, we implement a content-aware and active spatial filtering process. This process comprises two core components: a cross-attention block and a self-attention block. Specifically, collaborators utilize the correlation map as the query and local BEV feature as the key and value, to generate an updated content-aware spatial feature via the cross-attention block. Meanwhile, we employ a self-attention block on the BEV feature to capture its intrinsic dependencies within the feature space. Subsequently, the outputs of these two blocks are concatenated and processed by a lightweight MLP network. It performs channel reduction and outputs a self-activation mask map. Within this mask, cells exhibiting top-k activation values are identified as spatially complementary and perceptually significant via τ_k, yielding binary mask $\mathcal{M} \in \mathbb{R}^{H\times W}$. The whole process can be formulated as Equation 2.

$$\begin{aligned} G_i' &= [\mathrm{CrossAtt}(G_i, H_i) \parallel \mathrm{SelfAtt}(H_i)] \\ \mathcal{M}_i &= \tau_k(MLP(G_i' \in \mathbb{R}^{C_{emb}\times H\times W})) \end{aligned} \tag{2}$$

Spatial Refinement To further enhance the quality of shared features, a deformable attention block is employed to concurrently expand receptive fields and alleviate BEV spatial inconsistency induced by monocular depth uncertainty. In specific, the locations (x, y) of selected cells are regarded as reference points. Through a set of learned offsets, this mechanism facilitates spatial interactions with neighbor and related cells. This context re-sampling process counteracts depth-induced geometric inconsistency and refines spatial correlations when isolated indexing. Furthermore, this mechanism implicitly encodes spatial cues into the channels of selected cells, thus, an expanded receptive field is achieved without additional bandwidth consumption. This deformable attention is formulated as Equation 3, where $P_i = \{(x, y) \mid \mathcal{M}_i(x, y) = 1\}$.

$$H_i^{refine} = \mathrm{DeformAtt}(G_i', H_i, P_i) \tag{3}$$

Here, spatial-complementary and perceptual-significant features $H_{i\to ego}$ is derived by a simple selection operator.

Multi-Scale Pyramidal Prediction Network

Proposed multi-scale pyramidal prediction network ingests preserved historical BEV features $\mathbf{F}_0$ as unique input and outputting predicted semantic features at target resolution.

The architecture employs a top-down pyramidal hierarchy comprising L levels for feature prediction and aggregation. In specific, a bilinear-convolution upsampling module $u(\cdot)$ progressively refines high-dimensional features $\mathbf{F}_0$ and initialized queries $\mathbf{Q}_0$ across pyramid levels. For each level l, STDA predictor takes value $\mathbf{F}_l$ and query $\mathbf{Q}_l$ as input, producing prediction map $\mathbf{P}_l$. The prediction from the higher pyramid level $\mathbf{P}_l$ is upsampled by $u(\cdot)$ and concatenated with $\mathbf{P}_{l+1}$ along channel dimensions. Consequently, the concatenated features are aggregated by a learnable fusion block φ composed of a depth-wise convolution and dual point-wise convolution layers, enabling adaptive weighting of cross-scale cues. The whole process is formulated as:

$$\begin{aligned} \hat{\mathbf{P}}_{l+1} &= \varphi(u(\mathbf{P}_l)||\mathbf{P}_{l+1}), \quad l = 0, 1, \ldots, L \\ \mathbf{P}_{l+1} &= \mathrm{STDA}(u(\mathbf{F}_l), u(\mathbf{Q}_l)) \end{aligned} \tag{4}$$

In this hierarchical approach, coarse-scale predictions capture global structures while finer resolutions preserve local details. Through efficient aggregations, multi-scale information are fully leveraged to enhance prediction robustness.

Spatial-Temporal Deformable Attention (STDA)

STDA module is the core component in our prediction network. Inspired by (Li et al. 2022), we extend BEV deformable attention to spatiotemporal prediction contexts, by jointly modeling spatiotemporal correlations across sequential BEV features via learnable 3D offsets. This method dynamically attends to evolution patterns across both domains, effectively encoding time-dependent feature variations at each grid cell from the ego-centric perspective.

Query Initialization First, we predefine a group of grid-shaped learnable parameters for query initialization. Integrated with positional and temporal encoding, we can derive the initial query $\mathbf{Q} \in \mathbb{R}^{T_p\times C_{emb}\times H\times W}$, where T_p future time-steps are involved and each Q_t is with initial embedding dimension of C_{emb}.

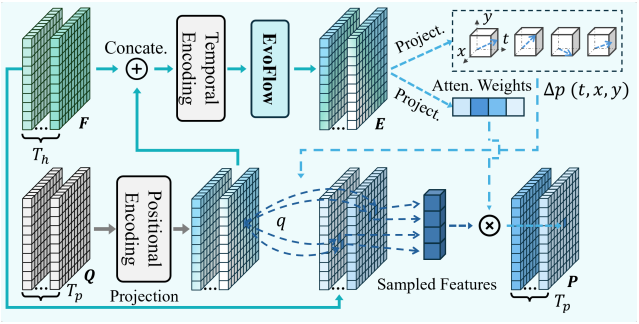


Figure 4: The architecture of proposed STDA module that models spatiotemporal correlations via 3D deformable attention, and an example query cell q is illustrated.

3D Offsets Learning To model holistic scene dynamics, we design a multi-layer convolutional network EvoFlow to capture spatiotemporal transition patterns and scene evolution cues across sequential features, inspired by flow field methods (Hui, Tang, and Loy 2018). Given historical features $\mathbf{F} \in \mathbb{R}^{T_h \times C \times H \times W}$, and concatenated with query $\mathbf{Q}$, EvoFlow will output scene dynamic representations $\mathbf{E}$, where each element E represents spatiotemporal correlations between historical and target predicted frames. Next, 3D offsets can be learned through a MLP and sigmoid activation. The whole process is formulated as:

$$\Delta p(t, x, y) = \sigma(\text{MLP}(\text{EvoFlow}(TE(\mathbf{F}||PE(\mathbf{Q}))))) \quad (5)$$

Where PE, TE denotes positional encoding and temporal encoding, and σ is Sigmoid activation.

3D Deformable Attention Subsequently, a group of 3D reference points $N_{ref} \in \mathbb{R}^{T_p \times H \times W \times k \times 3}$ are sampled, where k is the number of sampled points at each query cell $q \in \mathbf{Q}$ located at (t, h, w). Based upon Equation 1, our STDA module can be formulated as:

$$\text{STDA}(\mathbf{Q}, \mathbf{F}, N_{ref}) = \text{DeformAtten}(\mathbf{Q}, \mathbf{F}, N_{ref}) \quad (6)$$

Within this process, the reference points are updated by learned offsets that represents region of interest on historical features for each query cell q. Therefore, the queries can dynamically attend to relevant value features through its associated offsets, and then adaptively aggregate these features through learned attention weights. Eventually, STDA outputs multi-step future predictions $\mathbf{P}_p = [P_{t_0+1}, \ldots, P_{t_0+T_p}] \in \mathbb{R}^{T_p \times C \times H \times W}$.

Experiments

Experiment Setup

Dataset We employ multi-class BEV semantic segmentation and scene prediction as target tasks. The original OPV2V dataset (Xu et al. 2022a) provides only three semantic labels, thus, we evaluate the proposed methods on the enhanced version released in (Song et al. 2024), built upon OpenCDA (Xu et al. 2021) framework and CARLA simulator (Dosovitskiy et al. 2017). It totally records 11,464 frames and each frame includes two-to-seven connected vehicles, equipped with four cameras (800 × 600 resolution). Following a similar data processing pipeline, we generate 10-class ground-truth labels via dense point clouds aggregated from multiple surrounding LiDAR sensors. BEV cells that cannot be reasoned due to occlusion are assigned as 'unknown'. Furthermore, the labels of 'Vehicle' and 'Lane' will be completed and improved leveraging original dataset.

Implement Details

It is assumed that metadata sharing is well-synchronized and a random CAV is selected to serve as the ego agent among up to four CAVs. During experiments, we employ pretrained ResNet34 (He et al. 2016) as the backbone for image feature extraction, and employ FAX to perform image-to-BEV transformation following (Xu et al. 2022a). The evaluation BEV range is $[100m \times 100m]$ with a grid resolution of $0.39m$, which substantially exceeds typical range of individual perception systems. We first train the frame-wise collaborative BEV segmentation network for 48 epochs. Subsequently, with the pre-trained network parameters frozen, we extend training for 48 epochs on the auxiliary multi-frame forecasting branch. For the prediction task, $3.2s$ historical information are leveraged to forecast future $2s$. We employ a hybrid loss function for both semantic tasks, which is a weighted combination of Softmax Focal Loss, Lovász-Softmax Loss, and Dice Loss. Meanwhile, safety-critical categories Vehicle and Lane receive additional supervision through cross-entropy Loss terms. The whole model is trained with AdamW optimizer (Loshchilov and Hutter 2019) and cosine annealing learning rate scheduler (Loshchilov and Hutter 2017).

Evaluation Metrics We adopt the Intersection over Union (IoU) and mean IoU between predicted semantic maps and ground truth to evaluate the performances. Meanwhile, the transmission volumes and communication phases are leveraged to quantify the communication efficiency. Here the transmission volume is measured in standardized to Kibibyte (KiB) and log-scale base-2 bytes.

Quantitative Evaluation

Frame-Wise Perception To validate the effectiveness of the proposed ZeroR module, we conduct comparative experiments with four baselines leveraging fully sharing, including V2VNet (Wang et al. 2020), V2XViT (Xu et al. 2022b), AttenFuse (Xu et al. 2022c), and CoBEVT (Xu et al. 2022a). Our ZeroR module is deployed as a plug-in component integrated into CoBEVT baseline. Table 1 presents the performance comparison results. To provide an intuitive assessment of overall performance, we report average rank based on three prioritized evaluation metrics for holistic ranking, including mean IoU, communication protocol, and transmission volume. It is observed that our ZeroR module achieves improvement of 5.6% on mIoU, while reducing the transmission volume by half. Meanwhile, the integration with CoBEVT substantially enhances its overall ranking

Methods	Evaluation IoU										Commu.		**Rank**
	Build.	Fence	Pole	Road	Side.	Veg.	Terr.	**Lane**	**Veh.**	**mIoU***	**Protocol***	**KiB***	
V2VNet	7.8	18.7	2.7	77.2	39.0	25.6	43.3	42.8	50.4	34.2	-	512.0	**9**
V2XViT	8.9	20.0	3.1	78.4	40.4	28.3	47.3	45.1	50.4	35.8	-	256.0	**4**
AttenFuse	9.0	18.6	3.9	76.6	38.9	27.4	47.4	43.6	47.3	34.7	-	256.0	**7**
CoBEVT	8.9	17.2	4.5	75.9	38.2	25.7	44.3	43.1	49.7	34.2	-	256.0	**8**
+ SCAC	9.4	19.4	4.6	77.3	39.3	27.4	46.3	42.0	50.0	35.1	Dual	137.3	**6**
+ SCMF	9.1	19.4	4.1	76.1	39.3	26.5	47.1	43.3	49.4	34.9	Dual	132.3	**6**
+ EntropyCS	9.6	20.7	3.6	77.4	40.9	30.0	47.5	45.8	43.7	35.5	Dual	132.0	**3**
+ Max Pool	9.5	18.3	1.7	79.2	40.0	28.8	44.4	44.5	46.0	34.7	Single	128.0	**5**
+ FMS	9.6	19.8	2.9	77.0	39.5	29.0	47.8	45.8	47.4	35.4	Single	128.0	**2**
+ ZeroR (Ours)	9.0	20.0	5.8	78.7	40.3	28.9	46.8	45.2	50.2	36.1	Single	128.0	**1**

Table 1: BEV semantic segmentation performances comparison results. We report **Average Rank** of three prioritized evaluation metrics that are marked with $\star$ for holistic ranking. Spatial filtering methods are performed based on CoBEVT baseline.

Methods	mIoU					Lane IoU					Vehicle IoU				
	$0.4s$	$0.8s$	$1.2s$	$1.6s$	$2.0s$	$0.4s$	$0.8s$	$1.2s$	$1.6s$	$2.0s$	$0.4s$	$0.8s$	$1.2s$	$1.6s$	$2.0s$
ConvLSTM	27.6	25.9	24.8	23.9	22.8	<u>31.8</u>	30.2	28.9	27.8	26.4	23.5	17.0	13.7	11.9	10.4
PredRNN	27.2	25.7	24.5	23.4	22.5	30.6	29.1	27.8	26.5	25.3	25.1	18.3	13.6	10.6	9.6
Fiery*	27.5	25.8	24.2	23.3	22.4	29.9	29.4	28.5	27.8	26.9	<u>27.8</u>	20.1	16.1	14.2	12.8
ST-P3*	<u>27.8</u>	<u>26.5</u>	<u>25.7</u>	<u>25.0</u>	<u>24.3</u>	31.5	<u>30.4</u>	**29.7**	**29.1**	**28.4**	25.1	<u>20.9</u>	<u>17.9</u>	<u>15.8</u>	**14.5**
ZeRCP (Ours)	**29.5**	**27.3**	**26.1**	**25.4**	**24.6**	**32.8**	**31.2**	**29.7**	<u>28.9</u>	<u>28.2</u>	**29.3**	**21.9**	**18.5**	**16.5**	<u>14.1</u>

Table 2: BEV semantic scene prediction performances compared to other spatiotemporal prediction methods. **Boldface** and <u>underline</u> highlight the best and second-best performance. * denotes we re-implement the temporal modules.

from Rank 8 to Rank 1, even surpassing the top-performing V2XViT model under fully sharing paradigm.

Furthermore, to benchmark against SOTA communication efficient approaches, we re-implement their spatial compression modules and also adapt them to the same baseline under a close bandwidth budget. All other components and configurations are maintained identical to ensure a fair comparison. These modules include SCAC in Where2comm (Hu et al. 2022b), SCMF in How2comm (Yang et al. 2023a), and EntropyCS in UMC (Wang et al. 2023b) that follows dual-phase protocol, as well as Max-Pool in CORE (Wang et al. 2023a) and FMS in ERMVP (Zhang et al. 2024) that follows single-phase protocol. As presented in Table 1, our methods outperforms previous methods on overall ranking. In particular, the SOTA performances of mIoU is improved by 1.7% and 2.0% compared to dual-phase and single-phase protocols respectively. Another noteworthy improvement is observed for the Veh class, a critical category in autonomous driving scenarios, where ZeroR surpasses previous SOTA methods by 5.9% under single-phase protocol.

Future Scene Prediction Next, to validate the effectiveness of the proposed prediction network, we conduct comparative experiments against two typical spatiotemporal prediction baselines ConvLSTM (Shi et al. 2015) and PredRNN (Wang et al. 2017). Meanwhile, we adopt two methods that serve the same function within end-to-end autonomous driving frameworks Fiery (Hu et al. 2021) and ST-P3 (Hu et al. 2022a) for further evaluations. We re-implement them by substituting our prediction network within the whole framework, with all other components and configurations maintaining identical for fair comparison.

Quantitative results presented in Table 2 demonstrate significant performance advantages based on mIoU and two critical semantic categories. Our prediction method outperforms baselines by 1.8 mIoU, 1.7 Lane IoU, and 4.7 Veh IoU when averaged across all prediction horizons. Moreover, compared to probabilistic prediction approaches in end-to-end AD frameworks, our method achieves superior mIoU and Veh IoU with cumulative improvements of 3.6 and 6.1 points respectively, while maintaining comparable Lane IoU performance. These results substantiate STDA's capability to effectively model spatiotemporal correlations, particularly excelling in short-term predictions and dynamic patterns, while achieving long-term performance parity with recurrent architectures.

Ablation Study

Bandwidth Analysis Figure 5 illustrates the performance comparison across mIoU and two critical object categories under varying communication bandwidth conditions that are systematically modulated through distinct spatial filtering ratios. Key quantitative insights demonstrate that our method consistently surpasses prior spatial filtering approaches in both mIoU and Vehicle IoU across most of bandwidth conditions. Moreover, ZeroR attains better performances as the SOTAs with less communication volume required, highlighting its superior trade-off between perception performance and communication efficiency.

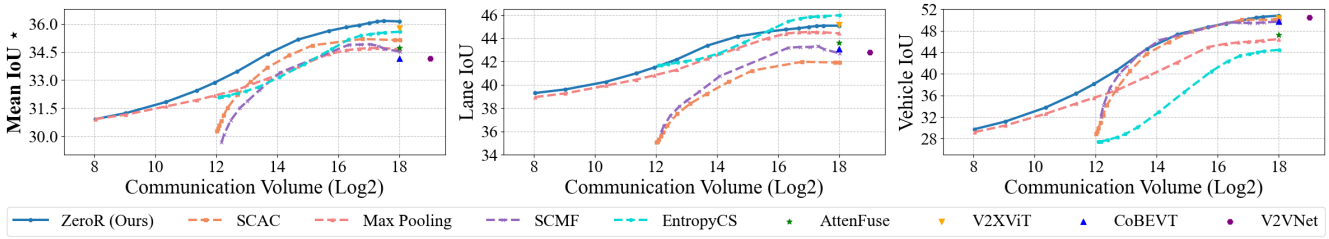


Figure 5: Performance comparison of distinct spatial filtering methods under varying communication volumes.

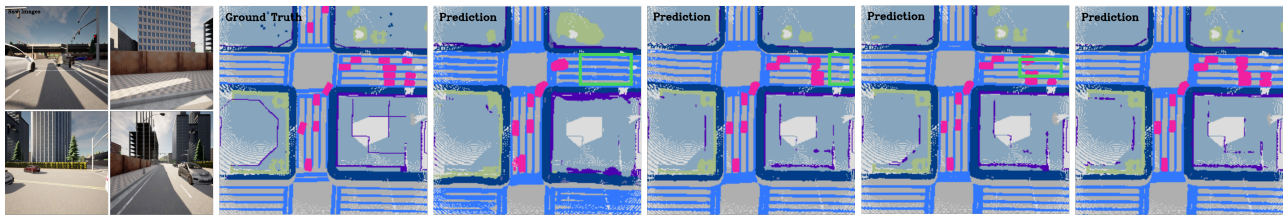


Figure 6: Qualitative comparison results. Our method ZeroR overcome most of the occlusions and accurately perceive distant objects, achieving better perceiving of the entire scene. Green boxes highlight advantages compared to other methods.

SCM	AFI	DSR	mIoU	Lane IoU	Veh. IoU
✓	✓	✓	**35.8**	**44.8**	**49.4**
	✓	✓	35.5	43.8	47.9
		✓	34.7	43.4	45.7
			33.9	42.5	43.7

Table 3: Ablation study results of proposed ZeroR with three core components: SCM, AFI and DSR.

Configurations	$0.4s$	$0.8s$	$1.2s$	$1.6s$	$2.0s$
$[4, 6, 8]$	29.5	27.3	26.1	25.4	24.6
$[2, 4, 6]$	29.2	27.3	26.1	25.2	24.3
$[4, 8, 10]$	29.7	27.4	26.2	25.4	24.3
$[6, 8, 10]$	29.8	27.5	26.2	25.3	24.2

Table 4: Ablation study results (mIoU) of offsets configuration in STDA module, with base configuration $[4, 6, 8]$.

Key Component Analysis Table 3 demonstrates the contributions of individual components within the proposed ZeroR module. In specific, we assess the impact of each component by sequentially removing (i) spatial correlation modeling (SCM), (ii) attention-based feature interaction (AFI), and (iii) deformable attention-based spatial refinement (DSR) from the whole module. Observed consistent performance degradation across all metrics validates the essential contribution of each component. Notably, integrating all three components yields substantial improvements of 3.8 mIoU, 2.5 Lane IoU, and 4.6 Veh IoU.

Number of Offset Sampling We investigate the impact of offset sampling numbers in STDA module, which reflects how many historical cells are referenced by each future query cell. In our base model, the number of offset samples per attention head is configured as $[4, 6, 8]$ across three pyramid scales. Typically, finer resolutions require more sample associations. Table 4 presents the performance sensitivity to different configurations. It presents that a modest number of samples achieves comparable prediction performance, validating the effectiveness of this method. Meanwhile, our selected configuration is carefully chosen to ensure neither insufficient nor redundant sampling.

Qualitative Evaluation

Figure 6 presents qualitative results of our method within a complex intersection scene. The visualization includes ego-vehicle's image inputs, ground truth, and comparative BEV segmentation outputs from different spatial filtering methods. All methods are performed under the same communication bandwidth condition. It is observed that collaboration techniques partially mitigate limitations of restricted sensing range and occlusions. Furthermore, our framework achieves more accurate perception of distant and occluded objects, and collectively delivers a more comprehensive understanding of the entire driving scene.

Conclusions

In this paper, we propose a communication-efficient framework ZeRCP that bridges collaborative perception with future scene prediction. We design a spatial filtering module ZeroR that eliminates reliance on requests while preserving inter-agent complementarity. Moreover, we extend frame-wise detection to multi-frame scene predictions through a multi-scale pyramidal prediction network anchored by STDA module. Extensive experiments prove the superiority of ZeroR module in bandwidth-constrained collaboration and efficacy of STDA in spatiotemporal modeling.

Acknowledgments

This work was partially supported by National Natural Science Foundation of China Grant 62373315, U24A20252, National Key Research and Development Program of China Grant 2024YFB4707603, and the Nansha Key Science and Technology Project 2023ZD006.

References

Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; and Koltun, V. 2017. CARLA: An open urban driving simulator. In *Conference on robot learning*, 1–16. PMLR.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, A.; Murez, Z.; Mohan, N.; Dudas, S.; Hawke, J.; Badrinarayanan, V.; Cipolla, R.; and Kendall, A. 2021. Fiery: Future instance prediction in bird's-eye view from surround monocular cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15273–15282.

Hu, S.; Chen, L.; Wu, P.; Li, H.; Yan, J.; and Tao, D. 2022a. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *European Conference on Computer Vision*, 533–549. Springer.

Hu, Y.; Fang, S.; Lei, Z.; Zhong, Y.; and Chen, S. 2022b. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. *Advances in neural information processing systems*, 35: 4874–4886.

Hu, Y.; Lu, Y.; Xu, R.; Xie, W.; Chen, S.; and Wang, Y. 2023. Collaboration helps camera overtake lidar in 3d detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9243–9252.

Hu, Y.; Peng, J.; Liu, S.; Ge, J.; Liu, S.; and Chen, S. 2024. Communication-efficient collaborative perception via information filling with codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15481–15490.

Hu, Y.; Shao, W.; Jiang, B.; Chen, J.; Chai, S.; Yang, Z.; Qian, J.; Zhou, H.; and Liu, Q. 2022c. Hope: Hierarchical spatial-temporal network for occupancy flow prediction. arXiv:2206.10118.

Huang, S.; Zhang, J.; Li, Y.; and Feng, C. 2024. Actformer: Scalable collaborative perception via active queries. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 14716–14723. IEEE.

Hui, T.-W.; Tang, X.; and Loy, C. C. 2018. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8981–8989.

Lang, A. H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; and Beijbom, O. 2019. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12697–12705.

Li, B.; Jin, X.; Wang, J.; Shi, Y.; Sun, Y.; Wang, X.; Ma, Z.; Xie, B.; Ma, C.; Yang, X.; et al. 2025. OccScene: Semantic occupancy-based cross-task mutual learning for 3D scene generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Li, J.; Xu, R.; Liu, X.; Ma, J.; Chi, Z.; Ma, J.; and Yu, H. 2023. Learning for vehicle-to-vehicle cooperative perception under lossy communication. *IEEE Transactions on Intelligent Vehicles*, 8(4): 2650–2660.

Li, Y.; Ren, S.; Wu, P.; Chen, S.; Feng, C.; and Zhang, W. 2021. Learning distilled collaboration graph for multi-agent perception. *Advances in Neural Information Processing Systems*, 34: 29541–29552.

Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Qiao, Y.; and Dai, J. 2022. BEVFormer: Learning Bird's-Eye-View Representation from Multi-camera Images via Spatiotemporal Transformers. In *European Conference on Computer Vision (ECCV)*, 1–18.

Liu, H.; Huang, Z.; and Lv, C. 2023. Multi-modal hierarchical transformer for occupancy flow field prediction in autonomous driving. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 1449–1455. IEEE.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Loshchilov, I.; and Hutter, F. 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. In *International Conference on Learning Representations*.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Ren, S.; Chen, S.; and Zhang, W. 2024. Collaborative joint perception and prediction for autonomous driving. *Sensors*, 24(19): 6263.

Shi, X.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; and Woo, W.-c. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28.

Song, R.; Liang, C.; Cao, H.; Yan, Z.; Zimmer, W.; Gross, M.; Festag, A.; and Knoll, A. 2024. Collaborative Semantic Occupancy Prediction with Hybrid Feature Fusion in Connected Automated Vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17996–18006.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, B.; Zhang, L.; Wang, Z.; Zhao, Y.; and Zhou, T. 2023a. Core: Cooperative reconstruction for multi-agent perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8710–8720.

Wang, T.; Chen, G.; Chen, K.; Liu, Z.; Zhang, B.; Knoll, A.; and Jiang, C. 2023b. Umc: A unified bandwidth-efficient

and multi-resolution based collaborative perception framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8187–8196.

Wang, T.-H.; Manivasagam, S.; Liang, M.; Yang, B.; Zeng, W.; and Urtasun, R. 2020. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, 605–621. Springer.

Wang, Y.; Long, M.; Wang, J.; Gao, Z.; and Yu, P. S. 2017. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. *Advances in neural information processing systems*, 30.

Wang, Z.; Wang, Y.; Wu, Z.; Ma, H.; Li, Z.; Qiu, H.; and Li, J. 2025. Cmp: Cooperative motion prediction with multi-agent communication. *IEEE Robotics and Automation Letters*.

Wei, S.; Wei, Y.; Hu, Y.; Lu, Y.; Zhong, Y.; Chen, S.; and Zhang, Y. 2023. Asynchrony-robust collaborative perception via bird's eye view flow. *Advances in Neural Information Processing Systems*, 36: 28462–28477.

Xu, R.; Chen, C.-J.; Tu, Z.; and Yang, M.-H. 2024. V2x-vitv2: Improved vision transformers for vehicle-to-everything cooperative perception. *IEEE transactions on pattern analysis and machine intelligence*.

Xu, R.; Guo, Y.; Han, X.; Xia, X.; Xiang, H.; and Ma, J. 2021. OpenCDA: an open cooperative driving automation framework integrated with co-simulation. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, 1155–1162. IEEE.

Xu, R.; Tu, Z.; Xiang, H.; Shao, W.; Zhou, B.; and Ma, J. 2022a. CoBEVT: Cooperative Bird's Eye View Semantic Segmentation with Sparse Transformers. In *Conference on Robot Learning (CoRL)*.

Xu, R.; Xiang, H.; Tu, Z.; Xia, X.; Yang, M.-H.; and Ma, J. 2022b. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *European conference on computer vision*, 107–124. Springer.

Xu, R.; Xiang, H.; Xia, X.; Han, X.; Li, J.; and Ma, J. 2022c. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *2022 International Conference on Robotics and Automation (ICRA)*, 2583–2589. IEEE.

Yang, D.; Yang, K.; Wang, Y.; Liu, J.; Xu, Z.; Yin, R.; Zhai, P.; and Zhang, L. 2023a. How2comm: Communication-efficient and collaboration-pragmatic multi-agent perception. *Advances in Neural Information Processing Systems*, 36: 25151–25164.

Yang, K.; Yang, D.; Zhang, J.; Li, M.; Liu, Y.; Liu, J.; Wang, H.; Sun, P.; and Song, L. 2023b. Spatio-temporal domain awareness for multi-agent collaborative perception. In *Proceedings of the IEEE/CVF international conference on computer vision*, 23383–23392.

Yang, K.; Yang, D.; Zhang, J.; Wang, H.; Sun, P.; and Song, L. 2023c. What2comm: Towards communication-efficient collaborative perception via feature decoupling. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7686–7695.

Yu, H.; Luo, Y.; Shu, M.; Huo, Y.; Yang, Z.; Shi, Y.; Guo, Z.; Li, H.; Hu, X.; Yuan, J.; et al. 2022. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 21361–21370.

Yu, H.; Tang, Y.; Xie, E.; Mao, J.; Luo, P.; and Nie, Z. 2023. Flow-based feature fusion for vehicle-infrastructure cooperative 3d object detection. *Advances in Neural Information Processing Systems*, 36: 34493–34503.

Yuan, X.; Kortylewski, A.; Sun, Y.; and Yuille, A. 2021. Robust instance segmentation through reasoning about multi-object occlusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11141–11150.

Zhang, J.; Yang, K.; Wang, Y.; Wang, H.; Sun, P.; and Song, L. 2024. Ermvp: Communication-efficient and collaboration-robust multi-vehicle perception in challenging environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12575–12584.

Zhao, B.; ZHANG, W.; and Zou, Z. 2023. BM2CP: Efficient Collaborative Perception with LiDAR-Camera Modalities. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, 1022–1035. PMLR.

Zhou, B.; and Krähenbühl, P. 2022. Cross-view transformers for real-time map-view semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13760–13769.

Zhou, Z.; Xiang, H.; Zheng, Z.; Zhao, S. Z.; Lei, M.; Zhang, Y.; Cai, T.; Liu, X.; Liu, J.; Bajji, M.; et al. 2024. V2xpnp: Vehicle-to-everything spatio-temporal fusion for multi-agent perception and prediction. arXiv:2412.01812.

Zyrianov, V.; Che, H.; Liu, Z.; and Wang, S. 2025. Lidardm: Generative lidar simulation in a generated world. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 6055–6062. IEEE.